Consumer Data Research Centre

An ESRC Data Investment

CACI

Personal Name Classification Using Collective Data

Hai Nguyen, Alexandros Alexiou and Alex Singleton
University of Liverpool

Impacts

- Adopts a range of alternative datasets such as social network website data, to mine linkages between family names.
- Demonstrates how unconventional data sources can be utilised for international names classifications.
- Aids the construction of name classification data across the globe independent of the availability of rich, full-coverage datasets that are scarcely available in many countries.

Project Background

Names can be used to reflect the cultural, ethnic and linguistic origins of individuals and much research has demonstrated how family names can be clustered into common origins using their associated given names. Deriving such classifications has a range of important applications, from understanding migration and population structures, to informing business planning, healthcare and local government decision-making. This research aims to address challenges in current name classification methods, such as their reliance on rich population data, by collecting and linking novel forms of data.

There are two important data types required to build a network of names: a set of full names and the frequency of such pairs. Current methodologies for automated classification of large names databases typically rely heavily on full-coverage data sources (such as an electoral register or a telephone directory), which creates significant limitations in countries where these data may be scarcely available. However, names can be collected from a range of alternative sources. Primarily, sources of interest are those that can be used in the identification of an individuals group, where a group can be either a country of origin, geographical region, ethnic group, or a religion. Yet, the majority of these data only contain either family names or given names and very often without frequency information.

In collaboration with CACI Ltd, this research presents a semi-supervised approach to classifying full names into ethnicity and nationality groups. The analysis uses a range of alternative data sources such as social network website data to mine linkages between family names and integrate them into an extensive international *name ranking* database. A case study is presented that evaluates the database using UK consumer full names.

Data and Methods

The solution adopted in this research was to mine names data (family names, given names, or full names) and their frequencies from multiple alternative data sources. Data were thus obtained from a social network website (approximately 180 million records) and various other sources such as the world names database, the electoral role, social media data (http://www.name-statistics.org) and national Censuses (UK 1881, US 1880). These datasets varied substantially in quality and quantity, for example, the number of records ranged from very limited (e.g. Greece) to very large (e.g. Poland). Larger databases typically presented more uncertainty than smaller ones, requiring database cleaning and verification processes.

Creating the Name Ranking Database

The methodology integrated these alternative data into an extensive names database. To achieve this, firstly, a family name graph was constructed, where each node of the graph represented a unique family name and a connection between two nodes represented at least one shared given name between two family names. Connections were weighted by frequencies of shared given names, indicating relationships between names within groups of individuals. Secondly, a clustering algorithm was applied which was sensitive to specific ethnic clusters emerging within other groups (for example, native Filipino family names within a Portuguese network). Thirdly, centrality measures were calculated for every graph, which was used to approximate the likelihood of each name belonging to a specific group.

The output of this process was a list of family names together with their associated non-negative centrality scores for each group. Similar centrality scores were calculated for given names and two tables were therefore produced, each containing three columns: family name/given name, group, and centrality score. This information was then classified into a singular group using a multi-stage algorithm that included identifying and classifying positive matches, performing basic pattern matching, removing hyphens/spaces and disambiguating records with multiple origin matches.

Consumer Data Research Centre

An ESRC Data Investment

CACI

Personal Name Classification Using Collective Data

Hai Nguyen, Alexandros Alexiou and Alex Singleton
University of Liverpool

Database Evaluation

To evaluate the name ranking database a UK case study was conducted, drawing comparisons between the database (consisting of 81 population groups) and a consumer register provided by CACI Ltd. This included the names and postcode-level addresses of over 54 million consumers in the UK, representing a high proportion of the adult population. The 2011 Census Country of Birth (CoB) dataset from the Office for National Statistics (ONS) provided a measure of the amount of people per country of birth at Lower Super Output Area (LSOA) and was used to assign country origins for both datasets. Group nationality estimates were obtained for each dataset and pairwise differences between ratios of the individual classes were then calculated at the LSOA level.

Key Findings

Pairwise differences are presented in Table 1. Results suggested that the name classification currently over or under-estimates groups, most significantly over-estimating the Irish population (8.38%) and underestimating the United Kingdom (3.84%).

Group	Average Difference	Standard Deviation	Upper Quantile	Lower Quantile
Europe, United Kingdom	-3.84	6.32	-0.90	-6.12
Europe, Ireland	8.38	10.46	12.28	3.17
Europe, EU Countries	1.98	4.60	4.29	0.46
Europe, Rest of	1.25	3.75	2.66	-0.45
Africa, North	-0.23	4.03	0.59	-1.13
Africa, Central and Western	0.44	3.23	0.63	-0.35
Africa, South and Eastern	-0.50	5.26	1.56	-1.57
Asia, Middle East	1.66	7.94	0.92	-0.34
Asia, Central	0.12	3.71	0.00	0.00
Asia, Southern	-0.20	3.62	0.45	-0.63
Asia, South-East	-1.67	3.55	-0.7	-2.52
Asia, Eastern	1.48	3.11	2.48	0.00

Table 1. Comparison between Name Classification and Census CoB

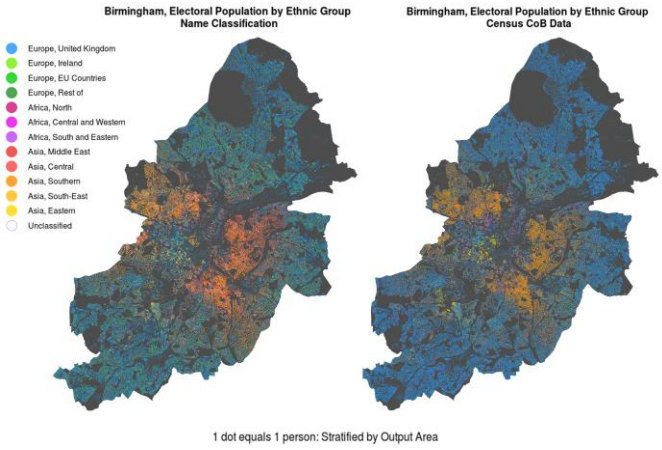

Figure 1. Dot-map presenting results from Name Classification vs. Census CoB data in Birmingham.

A finer-scale representation of ethnicity patterns can be seen using a dot-map of the Name Classification versus the Census CoB ratios within the Birmingham local authority (see Figure 1). Whilst these initial evaluations of the name ranking database suggest over or under estimations in some areas, results are promising at this preliminary stage. The researchers suggest that differences may be caused by second or third-generation immigrants that, although have British nationalities, their names indicate otherwise. For example, some minorities may adopt local (English) given names instead of the given names from their country of origin.

Current results and method development are preliminary, yet the names classification seems to respond well given the complexity of the issue. It is probable that over and under estimations are due to uncertainties in the data, of which the researchers suggest are likely a result of attempting to pool novel datasets that are fundamentally created as by-products of alternative agendas. However, these data offer promising examples of unconventional sources that can be utilised for international names classifications.

Development of this methodology hopes to address current challenges facing name classification research - primarily the lack of consistent data on an international scale. This may aid the construction of name classification data across the globe independent of the availability of rich, full-coverage datasets that are scarcely available in many countries. In addition, this research postulates the adoption of a semi-supervised approach, as opposed to traditional unsupervised methods, which can mean almost no a priori hypotheses can be used during the clustering process.

Future Directions

Future work will aim to address methodological issues such as the differentiation between groups of populations that have common names but different ethnic backgrounds (for example, Muslim populations). It will also continue to improve the accuracy of results by incorporating further novel datasets to the ranking database. The researchers hope that such progression demonstrates the ability to utilise and pool alternative datasets and contribute to the development of complex data mining techniques for name classification purposes.